Exercises

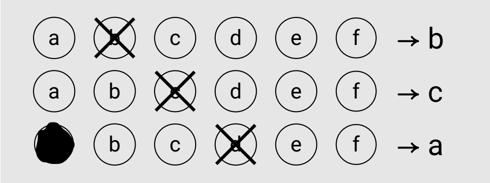
1	2	3	4	5

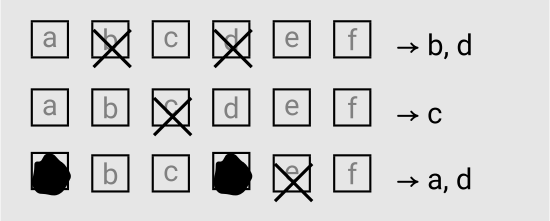
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Surname, First name

____

5SSD0 Bayesian machine learning and information processing
Final exam Q2





Fill in your answer(s) to the multiple-choice questions as shown above (circles = one correct answer, boxes = multiple correct answers possible).

Particular Ans on paper exam instructions

- Write in a black or blue pen.
- Hand in all pages. Do not remove the staple. If you remove it anyhow, check that you hand in all pages.

Dear student,

You're about to take an exam. Write down your name and your student ID at the appropriate places above. Make sure that you enter your student ID by fully coloring the appropriate boxes. On the examination attendance card, you fill in the PDF number. You can find the correct number on the top of the first page of your exam (e.g. 1234.pdf).

Please read the following information carefully:

Date exam: 2-Feb 2023
Start time 13.30
End time: 16.30 (+30 minutes for time extension students)

Number of questions: 5
Maximum number of points: 27
Answering style: multiple choice

Method of determining the final grade:

The maximal score for the written exam is $27/3 = 9.0$ points. The maximal score for the programming assignment is 1.0 point. These scores are added together for a score x =written + assignment. If (and only if) $5.0 \leq x \leq 5.4$, then I will take a look at your piazza activity. If you have been active and helpful on piazza (subjective measure, not up for discussion afterward), then I will give you a bonus and your score will become $x = 5.5$. The final score will be based on rounding towards whole integers, i.e., final score $= round(x)$. In other words, if $x >= 5.5$, you will pass the class.

Permitted examination aids

- Scrap paper (fully blank)
- (Formula sheet is included below)

Formula sheet

The **Gaussian distribution** is given by

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

The **multiplication of two multi-variate Gaussians** yields an (unnormalized) Gaussian:
$\mathcal{N}(x|\mu_a, \Sigma_a) \cdot \mathcal{N}(x|\mu_b, \Sigma_b) = \mathcal{N}(\mu_a|\,\mu_b, \Sigma_a + \Sigma_b) \cdot \mathcal{N}(x|\mu_c, \Sigma_c)$

where
$\Sigma_c^{-1} = \Sigma_a^{-1} + \Sigma_b^{-1}$
$\Sigma_c^{-1}\mu_c = \Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b$

The **Beta distribution** is given by
$\mathrm{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$,
where $\Gamma(\cdot)$ is the gamma function. The **mean of the Beta distribution** is given by $\mathrm{E}[x] = \frac{\alpha}{\alpha+\beta}$.

Important:

- You are only permitted to visit the toilets under supervision
- Examination scripts (fully completed examination paper, stating name, student number, etc.) must always be handed in
- The house rules must be observed during the examination
- The instructions of subject experts and invigilators must be followed
- Keep your workplace as clean as possible: put pencil case and breadbox away, limit snacks and drinks
- You are not permitted to share examination aids or lend them to each other
- Do not communicate with any other person by any means

During written examinations, the following actions will in any case be deemed to constitute fraud or attempted fraud:

- using another person's proof of identity/campus card (student identity card)
- having a mobile telephone or any other type of media-carrying device on your desk or in your clothes
- using, or attempting to use, unauthorized resources and aids, such as the internet, a mobile telephone, a smartwatch, smart glasses, etc.

- having any paper at hand other than that provided by TU/e, unless stated otherwise
- copying (in any form)
- visiting the toilet (or going outside) without permission or supervision

Generally, the final grade for this exam will be announced no later than fifteen working days after the date of this examination.

You can start the exam now, good luck!

1. Gaussian Model Predictions

A model m_1 for observations $x \in \mathbb{R}$ is specified as follows:

$$
\begin{aligned}
p(x|\mu, m_1) &= \mathcal{N}(x|\mu, 1) \\
p(\mu|m_1) &= \mathcal{N}(\mu|0, 1)
\end{aligned}
$$

(Consult the Formula sheet in the preamble of the exam for Gaussian distribution rules).

2p **1a** We make an observation $x = 1$. Determine the posterior $p(\mu|x = 1, m_1)$

ⓐ $\mathcal{N}(\mu|0, 0.5)$

ⓑ $\mathcal{N}(\mu|1, 2)$

● $\mathcal{N}(\mu|0.5, 0.5)$

ⓓ $\mathcal{N}(\mu|0.5, 1)$

2p **1b** Determine the evidence $p(x = 1|m_1)$ for model m_1, based on observation $x = 1$.

● $\mathcal{N}(1|0, 2)$

ⓑ $2/\sqrt{2\pi}$

ⓒ $\mathcal{N}(0|1, 1)$

ⓓ $1/\sqrt{2\pi}$

Consider a second model m_2, specified as

$$p(x|m_2) = \mathcal{N}(x|1, 1)$$

The model priors are given by $p(m_1) = 2/3$ and $p(m_2) = 1/3$.

1p **1c** Determine the probability $p(x = 2)$ by Bayesian model averaging over both m_1 and m_2.

ⓐ $\frac{2}{3\sqrt{2\pi}} + \frac{1}{3}\mathcal{N}(2|0, 1)$

ⓑ $\frac{1}{3}\mathcal{N}(2|1, 2) + \frac{1}{3\sqrt{2\pi}}$

● $\frac{2}{3}\mathcal{N}(2|0, 2) + \frac{1}{3}\mathcal{N}(2|1, 1)$

ⓓ $\frac{1}{3\sqrt{2\pi}} + \frac{1}{3}\mathcal{N}(2|1, 1)$

2. Classification

You have a machine that measures property x, the "transparency" of oils. You wish to discriminate between C_1 = 'olive oil' and C_2 = 'grape seed oil'. It is known that

$$p(x|C_1) = \begin{cases} -6(x-1)(x-2) & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x|C_2) = \begin{cases} 4-2x & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

The probability that x falls outside the interval $[1.0, 2.0]$ is zero. The prior class probabilities $p(C_1) = 0.4$ and $p(C_2) = 0.6$ are also known from experience.

2p **2a** Compute $p(C_1|x = 4/3)$.

ⓐ 2/3
ⓑ 3/4
ⓒ 3/5
● $4/10$

1p **2b** A "Bayes Classifier" is given by

$$\text{Decision} = \begin{cases} C_1 & \text{if } p(C_1|x) > p(C_2|x) \\ C_2 & \text{otherwise} \end{cases}$$

The Bayes classifier for this problem is given by:

ⓐ $\text{Decision} = \begin{cases} C_1 & \text{if } 4/3 < x \le 2 \\ C_2 & \text{otherwise} \end{cases}$

● $\text{Decision} = \begin{cases} C_1 & \text{if } 3/2 < x \le 2 \\ C_2 & \text{otherwise} \end{cases}$

ⓒ $\text{Decision} = \begin{cases} C_1 & \text{if } 5/3 < x \le 2 \\ C_2 & \text{otherwise} \end{cases}$

ⓓ $\text{Decision} = \begin{cases} C_1 & \text{if } 7/4 < x \le 2 \\ C_2 & \text{otherwise} \end{cases}$

2p **2c** Let the discrimination boundary be given by $x = a$. You make an observation and want to classify the observation. Work out the probability of making a wrong classification decision.

ⓐ $0.4\int_a^2 (4-2x)\mathrm{d}x + 0.6\int_1^a 6(1-x)(x-2)\mathrm{d}x$
● $0.6\int_a^2 (4-2x)\mathrm{d}x + 0.4\int_1^a 6(1-x)(x-2)\mathrm{d}x$
ⓒ $a/2$
ⓓ $\int_a^2 (4-2x)\mathrm{d}x + \int_1^a 6(1-x)(x-2)\mathrm{d}x$

3. Coin Toss Prediction

Consider a coin with outcomes

$$x_n = \begin{cases} 0 & \text{if tails is observed} \\ 1 & \text{if heads is observed} \end{cases}$$

We assume that the data-generating process is governed by a Bernoulli distribution,

$$p(x_n|\mu) = \mu^{x_n}(1-\mu)^{(1-x_n)}$$

and we assume a Beta distribution for the prior on μ:

$$p(\mu) = \mathrm{Beta}(\mu|\alpha = 3, \beta = 2)\,.$$

We throw the coin 5 times and observe outcomes $D = \{0, 1, 1, 0, 1\}$. (Consult the Formula sheet in the preamble of the exam for Beta distribution rules).

1p **3a** Work out the likelihood function $p(D|\mu)$ for μ.

● $p(D|\mu) = \mu^3(1-\mu)^2$

ⓑ $p(D|\mu) = \binom{5}{3} \cdot \mu^2(1-\mu)^3$

ⓒ $p(D|\mu) = \binom{5}{2} \cdot \mu^3(1-\mu)^2$

ⓓ $p(D|\mu) = \binom{3}{2} \cdot \mu^3(1-\mu)^2$

1p **3b** Compute the posterior distribution $p(\mu|D)$.

ⓐ $p(\mu|D) = \binom{5}{2} \cdot \mu^3(1-\mu)^2$

● $p(\mu|D) = \mathrm{Beta}(\mu|6, 4)$

ⓒ $p(\mu|D) = \mathrm{Beta}(\mu|5, 5)$

ⓓ $p(\mu|D) = \mu^3(1-\mu)^2 \cdot \mathrm{Beta}(\mu|\alpha = 3, \beta = 2)$

2p **3c** Compute the evidence $p(D)$.

ⓐ $p(D) = \frac{\Gamma(4)\Gamma(6)}{\Gamma(10)}$

● $p(D) = \frac{\Gamma(4)\Gamma(5)\Gamma(6)}{\Gamma(2)\Gamma(3)\Gamma(10)}$

ⓒ $p(D) = \frac{\Gamma(5)}{\Gamma(2)\Gamma(3)}$

ⓓ $p(D) = \frac{\Gamma(5)\Gamma(10)}{\Gamma(2)\Gamma(3)\Gamma(4)\Gamma(6)}$

1p **3d** Now compute the probability for throwing heads after the data set has been absorbed in the model.

ⓐ $p(x_{n+1} = 1|D) = \mathrm{Beta}(0.6|6,4)$

● $p(x_{n+1} = 1|D) = 0.6$

ⓒ $p(x_{n+1} = 1|D) = 0.7$

ⓓ $p(x_{n+1} = 1|D) = 0.4$

4. Miscellaneous comprehension

1p **4a** Consider a state-space model-based Active Inference agent that interacts with its world. Which of the following statements about the agent's computations is **most** consistent with Friston's FEP?

ⓐ Perception minimizes the *complexity of the states*.

ⓑ The agent infers actions by maximizing the *free energy in future states*.

ⓒ The agent infers actions by maximizing the *expected accuracy* in future states.

● The agent infers actions by minimizing the *expected free energy* in future states.

1p **4b** Overfitting occurs when a model fits too closely to a training data set, resulting in poor generalization. Why is a "Bayesian engineer" usually not very concerned about overfitting?

ⓐ Bayesian modeling aims to maximize (log-) model evidence, which decomposes as "training data fit minus model complexity". The complexity term prevents overfitting on the training data.

ⓑ Bayesian modeling uses a separate test data set to check the generalization properties of the model.

ⓒ Bayesian modeling uses probability theory to minimize the probability of overfitting as the training data set grows.

● Bayesian modeling aims to maximize (log-) model evidence, which decomposes as "training data fit minus entropy of the model parameters". The entropy term prevents overfitting on the training data.

1p **4c** Consider a data set $\{x_n|n = 1,2,\ldots,N\}$ with $x_n \in \mathbb{R}^M$ and a set of latent one-hot coded variables $z_n = (z_{n1}, z_{n2}, \ldots, z_{nK})$, i.e., $z_{nk} \in \{0,1\}$ and $\sum_{k=1}^K z_{nk} = 1$. Which of the following is a correct specification for a Gaussian Mixture Model?

ⓐ $p(x_n, z_n) = \prod_{k=1}^K \pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)$

● $p(x_n, z_n) = \prod_{k=1}^K (\pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k))^{z_{nk}}$

ⓒ $p(x_n, z_n) = \prod_{k=1}^K \pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}}$

ⓓ $p(x_n, z_n) = \prod_{k=1}^K (\pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k))^{z_n}$

1p **4d** A dark bag contains five red balls and seven green ones. Balls are not returned to the bag after each draw. If you know that on the last draw the ball was a green one, what is the probability of drawing a red ball on the first draw?

(a) 4/11 ● 5/11 (c) 5/12 (d) 6/11

2p **4e** Consider a (so-called Factor Analysis) model with specification

$$x_n = \Lambda z_n + v_n$$
$$z_n \sim \mathcal{N}(0, I)$$
$$v_n \sim \mathcal{N}(0, \Psi)$$

Furthermore, we assume that $\mathbb{E}[z_n v_n^T] = 0$. Evaluate $p(x_n)$.

(a) $p(x_n) \sim \mathcal{N}(0, \Lambda\Lambda^T + \Psi)$

● $p(x_n) \sim \mathcal{N}(0, \Lambda\Lambda^T + \Psi^T)$

(c) $p(x_n) \sim \mathcal{N}(1, \Lambda + \Psi)$

(d) $p(x_n) \sim \mathcal{N}(0, \Lambda + \Psi)$

1p **4f** Why can Variational Free Energy (VFE) minimization be interpreted as an approximation to Bayesian inference?

(a) VFE minimization is a model for Bayesian inference plus a little bit of Gaussian noise.

(b) VFE minimization minimizes the KL-divergence between the variational distribution and Bayesian evidence. Furthermore, the VFE itself is an upper bound on the Bayesian posterior distribution. Therefore, VFE minimization identifies approximations to both the posterior over latent variables and model evidence.

(c) VFE minimization minimizes Bayesian evidence by optimizing the variational posterior. Therefore, VFE minimization identifies approximations to both the posterior over latent variables and model evidence.

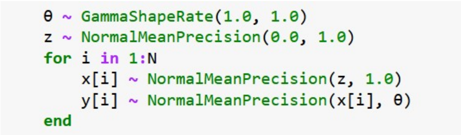
● VFE minimization minimizes the KL-divergence between the variational and Bayesian posterior distributions. Furthermore, the VFE itself is an upper bound to (negative log-)evidence. Therefore, VFE minimization identifies approximations to both the posterior over latent variables and model evidence.

5. Probabilistic Programming

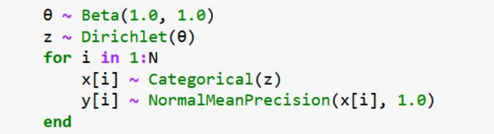
1p **5a** Consider the following probabilistic model:

$$p(y,x,z,\theta) = p(\theta)p(z)\prod_{i=1}^{N} p(y_i \mid x_i,\theta)p(x_i \mid z)$$

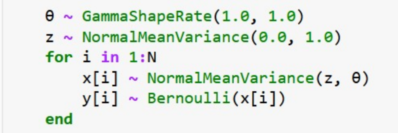
Which of the following specifications is consistent with the above model?

● 

```
θ ~ GammaShapeRate(1.0, 1.0)
z ~ NormalMeanPrecision(0.0, 1.0)
for i in 1:N
    x[i] ~ NormalMeanPrecision(z, 1.0)
    y[i] ~ NormalMeanPrecision(x[i], θ)
end
```

(b) 

```
θ ~ Beta(1.0, 1.0)
z ~ Dirichlet(θ)
for i in 1:N
    x[i] ~ Categorical(z)
    y[i] ~ NormalMeanPrecision(x[i], 1.0)
end
```

(c) 

```
θ ~ GammaShapeRate(1.0, 1.0)
z ~ NormalMeanVariance(0.0, 1.0)
for i in 1:N
    x[i] ~ NormalMeanVariance(z, θ)
    y[i] ~ Bernoulli(x[i])
end
```

(d) None of the code snippets are consistent with the model.

2p **5b** Suppose you are hired by an airplane company that wants to predict structural anomalies in their aircraft. They have decided to attach a thermometer to the landing gear, but are unsure how to handle all the variation in their measurements. Which of the following are reasonable assumptions for a probabilistic model? Multiple answers are possible.

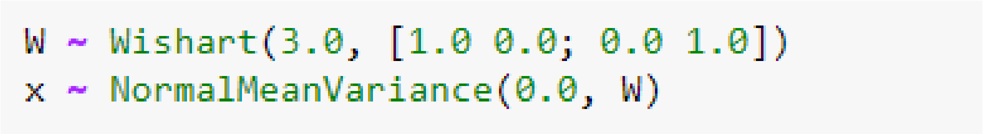
● The landing gear is either deployed (during landing) or retracted (during flight). We may assume that the data is generated according to a mixture: the first component (deployed) has high variance since temperature varies substantially across destinations and the second component (retracted) has low variance since the hold of the airplane is a controlled environment.

□ Temperature is measured continuously and varies slowly over time. If we assume that the state at time t is Gamma distributed and includes a state transition describing whether it becomes hotter or colder, then the state at time $t+1$ will be Gaussian distributed.

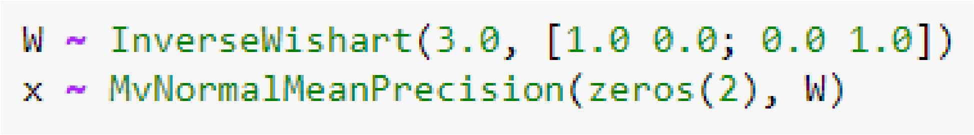
● We assume that the thermometer is not an ideal instrument and reflect this assumption by adding a zero-mean Gaussian-distributed measurement noise term in the model.

□ We assume that the measurements are *iid* (independent and identically distributed). We may therefore average the measurements to get rid of the noise.

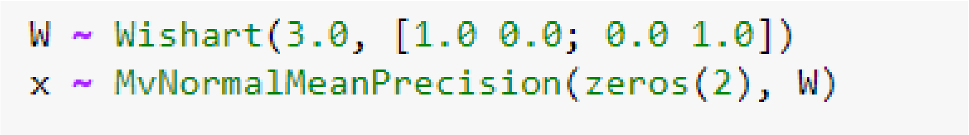
2p **5c** In which of the following code fragments is the distribution for W a conjugate prior? Multiple answers are possible.

□ 

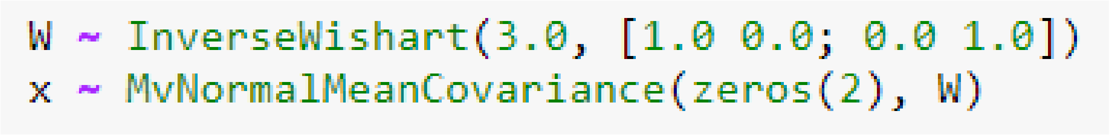
```
W ~ Wishart(3.0, [1.0 0.0; 0.0 1.0])
x ~ NormalMeanVariance(0.0, W)
```

□ 

```
W ~ InverseWishart(3.0, [1.0 0.0; 0.0 1.0])
x ~ MvNormalMeanPrecision(zeros(2), W)
```

● 

```
W ~ Wishart(3.0, [1.0 0.0; 0.0 1.0])
x ~ MvNormalMeanPrecision(zeros(2), W)
```

● 

```
W ~ InverseWishart(3.0, [1.0 0.0; 0.0 1.0])
x ~ MvNormalMeanCovariance(zeros(2), W)
```